



TITLE:

確率クラスタリング過程(離散数理モデルにおける最適組合せ構造)

AUTHOR(S):

渋谷, 政昭; 大和, 元

CITATION:

渋谷, 政昭 ...[et al]. 確率クラスタリング過程(離散数理モデルにおける最適組合せ構造). 数理解析研究所講究録 1993, 820: 33-46

ISSUE DATE:

1993-02

URL:

<http://hdl.handle.net/2433/83178>

RIGHT:

確率クラスタリング過程

慶大理工 渋谷 政昭 (MASAAKI SIBUYA)

鹿児島大理 大和 元 (HAJIME YAMATO)

1. 確率クラスタリング過程 (モデル A)

番号 $1, 2, \dots$ の付いた玉 B_1, B_2, \dots (各1個) を,
番号 $1, 2, \dots$ の付いた壺 U_1, U_2, \dots に適当に入れろ.
1個の壺には玉を何個でも入れることができる. 入り方を確
率規則を次の通りとする. 任意に $\alpha > 0$ は本稿で一貫して
用いるパラメータである.

1° B_1 を U_1 に入れる.

2° 以下 $n=1, 2, \dots$ に従って, “玉 B_1, \dots, B_n が
壺 U_1, \dots, U_u ($1 \leq u \leq n$) に入っており, U_j に入っ
ている玉の数が v_j ($1 \leq v_j, \sum_{j=1}^u v_j = n$)” の条件下で,

B_{n+1} は U_j に確率 $v_j/(\alpha+n)$ で入り, $1 \leq j \leq u$,

U_{u+1} に確率 $\alpha/(\alpha+n)$ で入る. \square

同じ壺に入る玉をクラスタとみるし, 上記の過程を‘確率クラ
スタリング過程’とよぶ. 本稿でこの4変種を考えるが, 上記

り、 Ξ と $\bar{\Xi}$ を区別する場合を 'モデル A ' と呼ぶ。 n 個の $\bar{\Xi}$ が入った状態を ' n 時点' と呼び、本稿では $\bar{\Xi}$ として $n < \infty$ における確率過程の分布を議論する。

Ξ B_n が入る Ξ の番号を X_n , $u = \max_{1 \leq j \leq n} X_j$,
 $A_{nj} = \{i: X_i = j\}$, $1 \leq j \leq u$, $V_{nj} = |A_{nj}|$ (集合 A_{nj} の基数) とし,

$$A^{(n)} = (A_{n1}, \dots, A_{nu}) \quad ; \quad V^{(n)} = (V_{n1}, \dots, V_{nu})$$

らどについて議論する。 $A^{(n)}$ の可能な実現値は,

$$\mathcal{A}_n = \{ a = (a_1, \dots, a_u), \quad a_j \neq \emptyset, \quad a_j \subset \mathbb{N}_n, \quad a_i \cap a_j = \emptyset \quad (i \neq j), \\ \bigcup_{j=1}^u a_j = \mathbb{N}_n, \quad \min \{ k \in \mathbb{N}_n, k \notin \bigcup_{i=1}^j a_i \} \in a_{j+1} \}$$

と、 $\mathbb{N}_n = \{1, \dots, n\}$ の、順序ある、制限付きの分割である。

また $V^{(n)}$ の可能な実現値は

$$\mathcal{V}_n = \{ v = (v_1, \dots, v_u) : v_j > 0, \sum_{j=1}^u v_j = n \}$$

と、 \mathbb{N}_n の、順序ある、自然数 n の分割である。

モデル A の前提条件を改めて書くと,

$$(1) \quad \begin{cases} P(X_1 = 1) = 1, \\ P(X_{n+1} = j \mid V^{(n)} = v) = \begin{cases} v_j / (\alpha + n), & 1 \leq j \leq u, \\ \alpha / (\alpha + n), & j = u+1, \end{cases} \end{cases}$$

ただし $v \in \mathcal{V}_n$, である。

さて, $A^{(n)}$ の確率分布が次の通りであることと, 帰納法により導くことができる:

命題 1. (1) の条件の下で

$$(2) \quad P(A^{(n)} = a) = \frac{\alpha^u}{\alpha^{[n]}} \prod_{j=1}^u (v_j - 1)!,$$

$$a = (a_1, \dots, a_u) \in \mathcal{A}_n, \quad v_j = |a_j|, \quad 1 \leq u \leq n,$$

$$\alpha^{[n]} = \alpha(\alpha+1) \cdots (\alpha+n-1)$$

である。□

$A^{(n)}$ の確率分布が $\{v_1, \dots, v_u\}$ に依存し、 v_j の順序にも a_j の要素にもよらないことを強調して置く。ただし $a \in \mathcal{A}_n$ の制約より a_j の要素は任意ではない。

命題 2. $C^{(n)} = C(C-1) \cdots (C-n+1)$ とする。

$$(3) \quad P(A_{n1} = a_1) = \frac{\alpha(v_1-1)!}{(\alpha+n-1)^{(v_1)}}, \quad v_1 = |a_1| = 1, \dots, n.$$

$$P(A_{n1}, A_{n2} = (a_1, a_2)) = \begin{cases} \frac{\alpha^2(v_1-1)!(v_2-1)!}{(\alpha+n-1)^{(v_1+v_2)}}, & v_j > 0, v_1+v_2 \leq n \text{ のとき}, \\ \frac{\alpha(n-1)!}{\alpha^{[n]}}, & v_1=n, v_2=0 \text{ のとき}, \end{cases}$$

...

$$P(A_{n1}, \dots, A_{nk} = (a_1, \dots, a_k)) = \begin{cases} \frac{\alpha^k \prod_{j=1}^k (v_j-1)!}{(\alpha+n-1)^{(\sum_{j=1}^k v_j)}}, & v_j > 0, \sum_{j=1}^k v_j \leq n \text{ のとき}, \\ \dots \\ \frac{\alpha^2(v_1-1)!(v_2-1)!}{\alpha^{[n]}}, & v_1 > 0, v_2 > 0, v_1+v_2=n, \\ & v_3 = \dots = v_k = 0 \text{ のとき}, \\ \frac{\alpha(n-1)!}{\alpha^{[n]}}, & v_1=n, v_2=\dots=v_k=0 \text{ のとき}, \end{cases}$$

ただし a_j は $(a_1, \dots, a_k, \Delta_n - \bigcup_{i=1}^k a_i) \in \mathcal{A}_n$ を満たすものに限る. \square

命題 2 から A_{n1}, A_{n2}, \dots についての条件付確率分布が得られる. たとえば

$$(4) \quad P(A_{n2} = a_2 \mid A_{n1} = a_1) = \frac{\alpha(v_2-1)!}{(\alpha+n-v_1-1)(v_2)}, \quad v_2 = 1, \dots, n-v_1.$$

この右辺は (3) の第 1 式で n を $n-v_1$ に変えたものである. このようにマルコフ性は $(X_n)_{n=1}^\infty$ のマルコフ性を反映したものであり, 一般的には次の命題が成り立つ.

命題 3.

$(A_{n1}, \dots, A_{nk}) = (a_1, \dots, a_k), \quad \sum_{j=1}^k |a_j| = v_0 < n$
 という条件の下での $(A_{n, k+1}, \dots, A_{nn})$ の確率分布は,
 $\Delta_n - \{1, \dots, v_0\}$ を Δ_{n-v_0} に変えれば, $A^{(n-v_0)}$ の確率分布に等しい. \square

確率クラスタリング過程に戻り, $(X_n)_{n=1}^\infty$ について命題 3 を考えると, $(X_{n_j} = 1)_{j=1}^\infty$ とする確率的部分系列を除いた残りが, $(X_n)_{n=1}^\infty$ と同じ構造をもつことになる. さらに

$X_{n_j} = 2$ の部分系列, $= 3$ の部分系列, \dots を除いても同様である.

2. 壺を区別できる場合 (モデル B)

確率クラスタリング過程で、壺を区別できる場合、 $A^{(n)}$ は単に \mathbb{N}_n の分割とみる事ができる。 $A^{(n)}$ の順序を無視したものを $A^{(n)*}$ で表わそう。その実現値は \mathbb{N}_n の分割の全体

$$A_n^* = \left\{ \{a_1, \dots, a_u\} : a_j \neq \emptyset, a_i \cup a_j = \emptyset, i \neq j, \right. \\ \left. \bigcup_{j=1}^u a_j = \mathbb{N}_n, 1 \leq u \leq n \right\}$$

である。部分集合の大きさを表わすにも、順序のある v を用いるのは不真当で、"順序統計量" $v_{(1)} \leq \dots \leq v_{(n)}$ を用いるか、あるいはこれと同等であるが、 $a_i = k$ とするに (あるいは v_i) の数 s_k , $1 \leq k \leq n$, を用いる。

$$S = (s_1, \dots, s_n), \quad s_k \geq 0, \quad \sum_{k=1}^n k s_k = n$$

を分割の '大きさの指標' と呼ぶ。 S は自然数 n の分割に他ならず、 $1^{s_1} 2^{s_2} \dots n^{s_n}$ という表示法も用いられている。

$u = \sum_{k=1}^n s_k$ は部分集合の数 (正の入った壺の数) である。

とはいっても、空の壺に正を入れるときに、もっとも小さい番号の壺に入れるために A_n は特別の構造をしており、分割 $\{a_1, \dots, a_u\}$ が与えられれば、その順序 (壺の番号) を復元できる: a_1 は 1 の入っている部分集合, a_2 は、 a_1 の要素外で最小のものが入っている部分集合, ... である。 U_1 は命題 2 でも見られるように、当然多くの正が入る

壺である。したがって、別のモデルとするのは必ずしも適切ではないが、壺の番号を無視した平象のみを問題にするときを‘モデル B’と呼ぶ。

命題 4. (1) の条件の下で、

$$(5) \quad P(A^{(n)*} = a^*) = \frac{\alpha^u}{\alpha^{[n]}} \prod_{j=1}^n ((j-1)!)^{s_j}, \quad a^* \in \mathcal{A}_n^*,$$

ただし $S = (s_1, \dots, s_n)$ は分割 a^* の下付きの指標, $u = \sum_{k=1}^n s_k$, である。□

(5) は (2) の書き換えにしか過ぎないが, $A^{(n)*}$ の分布が S にしか依存していることを改めて強調する。実際、この式の番号を $B_{\nu_1}, \dots, B_{\nu_n}$ ((ν_1, \dots, ν_n) は \mathbb{N}_n の任意の順列) に変えても (5) には関係しない。つまり, (5) は要素の置換に関して不変である。

この事実から、たとえば、“任意の $i, j \in \mathbb{N}_n$ が同じ部分集合に入る確率”は“1, 2 が同じ部分集合に入る確率”に等しく、したがって $A^{(n)}$ の分布について計算する必要はなく、

$$P(A^{(2)} = (\{1, 2\})) = \frac{1}{\alpha+1}$$

と求まる。同様の考察で、いろいろの平象の確率が計算できるが、一般的には次のようにまとめることができる。

$$\mathbb{N}_n = E_1 \cup E_2, \quad E_1 \cap E_2 = \emptyset, \quad |E_1| = m, \quad 0 < m < n$$

とし, $A^{(n)*} \cap E_i$, $i=1, 2, \dots$ $\{A_{n1} \cap E_1, \dots, A_{nn} \cap E_n\}$ から空集合を除いたものとする.

命題 5. b_i^* , $i=1, 2, \dots$ E_i の任意の分割とする.

(i) $a^* = b_1^* \cup b_2^*$ のときは

$$\begin{aligned} P(A^{(n)*} = a^* \mid A^{(n)*} \cap E_2 = b_2^*) \\ = P(A^{(n)*} \cap E_1 = b_1^* \mid A^{(n)*} \cap E_2 = b_2^*) = P(A^{(n)*} = b_1^{**}) \end{aligned}$$

ただし b_1^{**} は b_1^* の要素を Ω_m の要素に変えたものである.

(ii) 上の (i) と同じ記号で,

$$P(A^{(n)*} \cap E_1 = b_1^*) = P(A^{(n)*} = b_1^{**}).$$

たゞは $A^{(n)*}$ の任意の m 要素を取り上げ, それを無視したときの確率分布の分布である. \square

これらの事実は確率クラスタリング過程において本質的であり, 逆に分布 (2) を次のように特徴づけることができる.

命題 6. 任意の時点 n における \mathcal{A}_n 上の確率分布

$$P(A^{(n)} = a), \quad a \in \mathcal{A}_n, \quad \text{が}$$

(i) a の成分の基数の集合 $\{|a_1|, \dots, |a_n|\}$ だけに依存し,

(ii) $A_{n1} = \{1, \dots, k\}$, $1 \leq k \leq n-1$, の条件下での $(A_{n2},$

$\dots, A_{nn})$ の確率分布が, 要素 $\{k+1, \dots, n\} \in \Omega_{n-k}$

に変えれば, $A^{(n-k)}$ の確率分布に等しい.

このとき, $P(A^{(n)}=a)$ は (2) 式と与える. \square

命題 7.

命題 6 の条件 (ii) を次のように変えることができる:

(ii') N_n に属する任意の要素 j を $A^{(n)}$ から除いたときの確率分布が, 要素 j を $((T=1 \text{ ならば } 1))$ N_{n-1} に変えたとき, $A^{(n-1)}$ の分布に等しい. \square

3. \mathcal{A} を区別できる場合 (モデル C)

確率クラスタリング過程において \mathcal{A} が区別できず, 型に入る \mathcal{A} の数 v_i が観測可能であるとす. 時点 n で U_1, \dots, U_n に入る \mathcal{A} の数を $V^{(n)} = (V_{n1}, \dots, V_{nn})$ とする.

命題 8. (i) の条件の下で,

$V^{(n)} = (V_{n1}, \dots, V_{nn})$, $V_{nj} = |\{i: X_i = j\}|$, の確率分布は,

$$(b) \quad P(V^{(n)}=v) = \frac{\alpha^n}{\alpha^{[n]}} \frac{(n-1)!}{\prod_{j=1}^{u-1} (n - \sum_{i=1}^j v_i)},$$

$$v = (v_1, \dots, v_u) \in \mathcal{V}_n, \quad u = 1, \dots, n. \quad \square$$

この場合を 'モデル C' と呼ぶことにする. モデル A で単に v とする $a \in \mathcal{A}_n$ の数をかき入れる. 逆に, モデル C で v_1 個の \mathcal{A} に, 1 と $N_n \setminus 1$ から選んでランダムに番号を付け, 残りに逐次 $a \in \mathcal{A}_n$ の条件を満た

すよう にラングラムに付ければモデル A を得る.

モデル A の命題 2 に対応する $V^{(n)}$ の周辺分布は次の通りである.

命題 9.

$$(7) \quad P(V_{n1} = v) = \frac{(n-1)^{(v-1)} \alpha [n-v]}{(\alpha+1) [n-1]}, \quad v=1, \dots, n$$

あるいは

$$\begin{aligned} P(V_{n1}-1 = x) &= \binom{\alpha+n-x-2}{n-x-1} / \binom{\alpha+n-1}{n-1} \\ &= \binom{-1}{x} \binom{-\alpha}{n-1-x} / \binom{-\alpha-1}{n-1}, \quad x=0, \dots, n-1 \end{aligned}$$

これは負の超幾何分布 $\text{NegHge}(n-1; 1, \alpha)$, つまり 2 項分布 $\text{Bn}(n-1, p)$ の p が ベータ分布 $\text{BE}(1, \alpha)$ に従う ベータ・2 項分布である. \square

周辺分布 $P((V_{n1}, V_{n2}) = (v_1, v_2))$ なども同様の形になる. 条件分布も

$$P(V_{n2} = v_2 \mid V_{n1} = v_1) = \frac{(n-v_1-1)^{(v_2-1)} \alpha [n-v_1-v_2]}{(\alpha+1) [n-v_1-1]},$$

$$v_2 = 1, \dots, n-v_1.$$

これは, (7) の n を $n-v_1$ に変えたものと等しい. 一般に,

命題 10. $(V_{n1}, \dots, V_{nk}) = (v_1, \dots, v_k)$, $\sum_{j=1}^k v_j = v_0 < n$

という条件の下での $(V_{n,k+1}, \dots, V_{nu})$ の相対分布は,

(6) で $n \in n - v_0$ に変えたものである。□

これは命題 5 に相当するものである。命題 6 に相当するものは次の通りである。

命題 11. $V^{(n)}$ が (6) の確率分布に従うとする。

n 個の玉のうち 1 個をランダムに選んで除くとする。

もしも $v_k = 1$ の壺 U_k を選ぶと他の壺が空になるか、そのときには U_{k+1}, \dots, U_n の番号をひとつずつ前にずらす。

このときの玉の数 $V^{(n-1)}$ の分布は (6) で $n \in n-1$ に変えたものである。□

後述のように $n \rightarrow \infty$ の極限を考えるときにはモデル C が基本的とする。

4. 玉も壺も区別できる場合 (モデル D)

最後に、玉も壺も区別できず、部分集合の大きさの指標だけで観測できる場合を 'モデル D' とする。確率的な指標を $S^{(n)} = (S_{n1}, \dots, S_{nn})$ とする。その実現値の全体は自然数 n の分割の全体 J_n である。

命題 12. 確率クラスタリング過程の n 時点で、分割の大きさの指標 $S^{(n)}$ の確率分布は

$$(g) \quad P(S^{(n)} = s) = \frac{\alpha^4}{\alpha^{[n]}} \cdot \frac{n!}{\prod_{j=1}^n (j^{s_j} s_j!)} ,$$

$$S = (S_1, \dots, S_m) \in \mathcal{J}_m, \quad u = \sum_{j=1}^m S_j,$$

びある。□

分布 (8) はモデル B から、多項係数を印し、計算で直ちに得られるが、モデル C から導くことはそれほど単純ではない。逆にモデル D からモデル B を導くのも、単にラベルの番号付けをすればいいわけであるが、モデル C を導くには技巧が必要である。

命題 13. $S^{(n)}$ が (8) の分布に従うとする。

几个の要素のうち 1 つを等確率で選び、それが属している部分集合 (基数を V_1 とする) を除く。残りの $n - V_1$ 個の要素の 1 つを等確率で選び、それが属している部分集合 (基数を V_2 とする) を除く。... このとき (V_1, V_2, \dots, V_n) はモデル C の分布 (6) に従う。□

したがって n 個の要素の 1 つを等確率で選ぶ。その要素が属している部分集合全体を除いても、その要素 1 個を除いても (8) で n から除いた数だけ減らしただけで得られる。

5. 4つのモデル

4つのモデル A-D を適当に使い分けることにより、端対象の計算が容易となる。本質的の特徴は、正の番号の

置換不変性 (exchangeability) である。それぞれの確率分布を下の表にまとめておく。

確率クラスタリング過程の確率分布

置換不変性 \ 置換	置換あり	置換なし
置換あり	$A: P(A^{(n)} = a) = \frac{\alpha^n}{\alpha^{[n]}} \prod_{j=1}^n (v_j - 1)!$	$C: P(V^{(n)} = v) = \frac{\alpha^n}{\alpha^{[n]}} \frac{(n-1)!}{\prod_{j=1}^{u-1} (n - \sum_{i=1}^j v_i)}$
置換なし	$B: P(A^{(n)*} = a^*) = \frac{\alpha^n}{\alpha^{[n]}} \prod_{j=1}^n ((j-1)!)^{s_j}$	$D: P(S^{(n)} = s) = \frac{\alpha^n}{\alpha^{[n]}} \frac{n!}{\prod_{j=1}^n (j^{s_j} s_j!)}$

A: $a = (a_1, \dots, a_n) \in \mathcal{A}_n$. a_j : 置換 U_j に入っている正の番号の集合.
 $v_j = |a_j| > 0$.

B: $a^* = \{a_1, \dots, a_n\} \in \mathcal{A}_n^*$: \mathbb{N}_n の分割の全体
 $s = (s_1, \dots, s_n) \in \mathcal{I}_n$: n の分割の全体

C: $v = (v_1, \dots, v_n)$ A と同じ

D: $s = (s_1, \dots, s_n)$ B と同じ

(パラメータ)

いずれも α を含む部分は同じであり、正の入っている置換の数 u は十分統計量である。その確率分布は、

$$(9) \quad P(U_n = u) = \frac{[n]}{[u]} \frac{\alpha^u}{\alpha^{[n]}} \quad , \quad u = 1, \dots, n,$$

ただし $[n]$ は第 1 種スターリング数 (Knuth の記号)

であり、等式 $\alpha^{[n]} = \sum_{u=1}^n \binom{n}{u} \alpha^u$ によって定まる。

6. モデル C の極限分布.

モデル C において $V^{(n)}/n = (V_{n1}/n, \dots, V_{nu}/n)$

とし、 $n \rightarrow \infty$ とすると、無限次元単体上の確率分布が得られる。負の超幾何分布に従う V_{ni} は n で割った極限は、2項分布の比が確率に収束するから、ベータ分布であり、今の場合、この確率密度関数は

$$(10) \quad \alpha(1-u)^{\alpha-1} \cdot 1[0 < u < 1]$$

である。($1[\cdot]$ は平気の定義関数。) $V^{(n)}$ のマルコフ性 (命題 13) より、極限も同じ性質をもつ。

命題 14.

$(W_n)_{n=1}^{\infty}$ を (10) に従う $(0,1)$ 上の $\overset{\text{i.i.d.}}{\text{独立変数}}$ の系列とし、 $(U_n)_{n=1}^{\infty}$ を

$$U_1 = W_1, \quad U_n = W_n \prod_{j=1}^{n-1} (1 - W_j), \quad n=2, \dots$$

で定義する。 $V^{(n)}/n$ ($n \rightarrow \infty$) は $(U_n)_{n=1}^{\infty}$ に収束する。さらに (X_1, \dots, X_n) は $(U_n)_{n=1}^{\infty}$ からの確率標本である。 \square

$(U_n)_{n=1}^{\infty}$ は確率クラスタリング過程の諸性質を継承している。

7. おわりに.

本書は独創的でもなければ、最近集団遺伝学で得られている成果を、集団遺伝学の用語・概念を援けずに解説したものである。ただし p.12 の表のように 2×2 のモデルとして整理することにより体系的に、明快と云った自負している。文献は非常に多いが次の2つの総合報告に詳しい文献表がある。

- ・ Hoppe, F.M. (1987) The sampling theory of neutral alleles and an urn model in population genetics, J. Math. Biology 25 123-159.
- ・ Ewens, W.J. (1990) Population genetics theory — The past and the future, S. Lessard (ed.) Mathematical and Statistical Developments of Evolutionary Theory, 177-227, NATO Adv. Sci. Inst. Ser. C 299, Kluwer, Dordrecht, Holland.